From Regular to Context Free to Mildly Context Sensitive Tree Rewriting Systems:
The Path of Child Language Acquisition

Robert Frank
University of Delaware
Department of Linguistics
46 E. Delaware Avenue
Newark, DE 19716 USA
email: rfrank@cis.udel.edu

Abstract

Current syntactic theory limits the range of grammatical variation so severely that the logical problem
of grammar learning is trivial. Yet, children exhibit characteristic stages in syntactic development at
least through their sixth year. Rather than positing maturational delays, I suggest that acquisition
difficulties are the result of limitations in manipulating grammatical representations. I argue that the
genesis of complex sentences reflects increasing generative capacity in the systems generating structural
descriptions: conjoined clauses demand only a regular tree rewriting system; sentential embedding uses
a context-free tree substitution grammar; modification requires TAG, a mildly context-sensitive system.

I. Some current views of natural language syntax localize all cross-linguistic variation in a small set of
finite-valued parameters. This has the effect of eliminating certain in-principle learnability problems that
plagued earlier views of grammar. Since the range of possible hypotheses is restricted, the child will show an
inductive bias, that is, she will at times be driven to conclusions seemingly stronger than those warranted
by the input data. Further, learnability in the limit is guaranteed, since the number of possible grammars
is finite. Finally, the limited amount and simplicity of information to be acquired means that grammars
should be acquired quickly and easily. Unfortunately, this beautiful picture conflicts with what we know
from empirical studies of language acquisition: children exhibit characteristic developmental stages in their
acquisition of grammar. Consequently, if such a parametric view of syntactic variation is correct, children
must be held back in their attempts at syntactic learning by something other than the inherent difficulty of
the learning task.

I suggest that children's acquisitional difficulties result not from problems of grammatical acquisition per
se, but rather from their limited abilities in manipulating grammatical representations. In particular, I
argue that the sequence of certain stages in syntactic development can best be understood as a reflection
of ever increasing generative complexity of the underlying formal grammatical systems used by the child to
construct her tree structure representations.

II. It has been widely reported in empirical studies of language acquisition that different types of complex
(i.e. multi-clausal) sentences vary with respect to the point at which children first exhibit mastery of them.
Looking at the naturalistic production data of four children, Bloom et al. (1980) report that the productive
use of complex sentences involving conjunction consistently precedes that of sentences involving complemen-
tation which in turn precedes sentences involving relativization. This result is supported by experimental
study of children's comprehension. Tavakolian (1981) demonstrates that young children exhibit difficulty in
interpreting relative clauses. She argues that the interpretations that children do assign to such structures
result from their (incorrectly) imposing a conjoined clause analysis. I take this tendency to prefer conjunction
over relativization to be the same effect observed by Bloom and her colleagues. Further, Goodluck (1981),
Hsu et al. (1985), McDaniel and Cairns (1990) among others have found that children correctly interpret
"control constructions" in which the empty subject is within a complement clause, such as (1), at an earlier
age than cases where the empty subject is within an adverbial clause as in (2).

(1)   a. Cookie Monster tells Grover$_i$ [PRO$_i$ to jump over the fence]
        b. Grover$_i$ was told by Cookie Monster [PRO$_i$ to jump over the fence]
(2)   Cookie Monster$_i$ touches Grover [after PRO$_i$ jumping over the fence]

Let us suppose that a child's incorrect interpretations in (2) are the result of her inability to assign this
sentence a structural representation appropriate according to the adult grammar. If we assume that complex

sentences containing adverbial clause adjuncts are similar to sentences containing relative clauses in the relevant structural respects, i.e., they involve adjunction structures, this phenomenon can be seen as another instance of Bloom et al.'s sequence of complementation before relativization (now modification).

III. Thus far, we've seen that children's acquisition of complex constructions proceeds according to the sequence coordination < complementation < modification. Yet, we have not provided an explanation for why these should be so ordered. What I will now suggest is that these stages are ordered by the ever greater demands of generative capacity that they impose upon the formal tree rewriting system used to construct phrase structure representations.

Before proceeding with this, we will require a brief detour into defining a novel tree rewriting formalism. Developing a suggestion of Weir (1987) (though in a restricted fashion), let us define a schematic tree grammar (STG) as a 4-tuple $G = (V_N, V_T, S, I)$ where V_N is a finite set of non-terminals, V_T is a finite set of terminals, S is a distinguished non-terminal and I is a finite set of schematic initial trees. The set of schematic initial trees in an STG may be any finite set of finite tree structures whose frontier nodes are drawn from $V_T \cup V_N$ and whose internal nodes are drawn from V_N. Further, all nodes of schematic initial trees but the root may be annotated with the superscripts $+$ or $*$. The intuition behind the use of these superscripts is the same as their usage in regular expressions: each schematic tree represents an infinite set of trees, just as each regular expression represents an infinite set of strings. When a schematic tree contains a node N marked by $+$, the class of trees representd by this schematic tree includes trees containing 1 or more copies of the subtree dominated by N, each copy attached to N's parent. Similarly, schematic trees containing nodes marked $*$ correspond to those trees where this node appears 0 or more times. We formalize this as follows:

Definition 1 *A (possibly empty) sequence of trees $\langle \tau_1, \ldots \tau_k \rangle$ instantiates a schematic tree σ iff:*
1. if the root of σ is superscripted by $+$, the sequence of τ_i's is of length ≥ 1;
2. if the root of σ is not superscripted, the sequence of τ_i's is of length exactly 1;
3. for each τ_i, the root is labelled identically to the root of σ;
4. for each τ_i, the sequence of subtrees dominated by the children of the root of τ_i, $\langle \tau_i^1, \ldots \tau_i^n \rangle$, may be partitioned into a sequence of contiguous subsequences, $\langle \tau_i^1, \ldots \tau_i^j \rangle, \langle \tau_i^{j+1}, \ldots \tau_i^k \rangle \ldots \langle \tau_i^m, \ldots \tau_i^n \rangle$, so that these subsequences successively instantiate the subtrees dominated by the children of root of σ from left to right.

Derivations in an STG G do not directly utilize the schematic trees in I, but rather manipulate the trees which instantiate the trees in I. The only combinatory operation we allow in an STG is substitution. Application of substitution is however restricted so as to prevent the generation of recursive structures. The set of derivable trees from a grammar G, $D(G)$ is thus defined as follows:

Definition 2 *τ is derivable by G, $\tau \in D(G)$, iff:*
1. τ instantiates some $\sigma \in I$; or
2. τ is the result of substituting τ' into τ'' where $\tau', \tau'' \in D(G)$ and no non-terminal on the path from the root of τ'' to the site of subsitution appears in τ'.
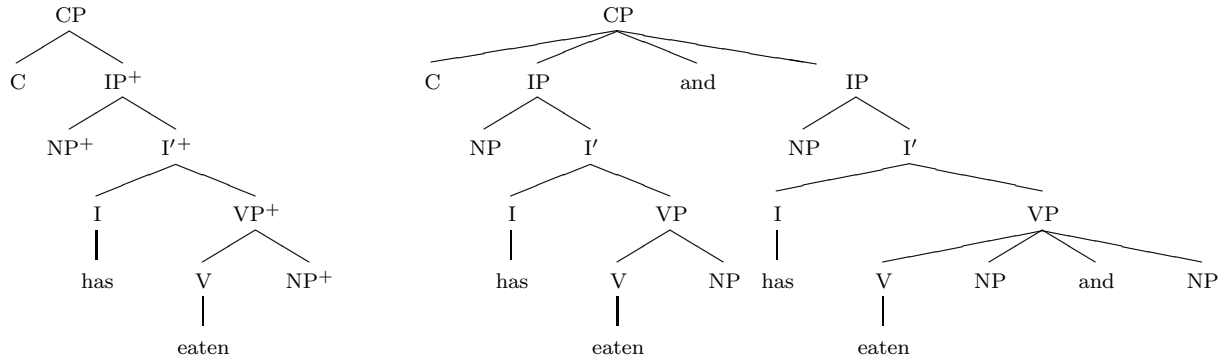
We finally define the set of trees and strings which are generated by an STG G as follows:

Definition 3 *The* tree set *of an STG G is the set of trees $T(G) = \{\tau | \tau \in D(G), \tau$ is rooted in S and the frontier of $\tau \in V_T^* \}$.*

Definition 4 *The* string language *of an STG G is $L(G) = \{w | w$ is the frontier of some $\tau \in T(G)\}$*

Two things concerning this tree rewriting system are relevant for our concerns. The first is that the string languages generated by STGs are all and only the regular languages. The second is that, in spite of its relatively weak formal power, it is nonetheless sufficiently expressive to express analyses of (certain cases of) coordination structures. If the schematic tree below on the left is in the grammar, the STG will generate the string *John has eaten an apple and Fred has eaten peaches and a candy bar*: the tree below on the right, which duplicates the IP root node once and duplicates the NP object, instantiates this schematic tree. We need only perform the relevant substitutions into the NP nodes to complete the derivation.[1]

---

[1]In the interest of space, we put aside certain complications concerning the insertion of conjunctions, the allowance for some degree of lexical variation within the internal structure of coordinated phrases. Such issues can be dealt with without increase in generative capacity.

Tree 1:
CP
- C
- IP$^+$
  - NP$^+$
  - I$'^+$
    - I — has
    - VP$^+$
      - V — eaten
      - NP$^+$

Tree 2:
CP
- C
- IP
  - NP
  - I$'$
    - I — has
    - VP
      - V — eaten
      - NP
- and
- IP
  - NP
  - I$'$
    - I — has
    - VP
      - V — eaten
      - NP
      - and
      - NP

This regular tree rewriting system does not allow for the generation of syntactic structures appropriate for sentential complementation. There is simply no way to produce the unbounded degree of embedding that is necessary from a linguistic standpoint. If, however, we move to the system of tree rewriting discussed by Schabes (1990) called Tree Substitution Grammars (TSG), complementation can be accommodated. A TSG consists of a finite set of (finite) elementary trees whose leaves may be either terminals or non-terminals. As before, derivations consist in substituting these elementary trees rooted in some category C into non-terminals at nodes along the frontiers of other elementary trees also labeled C, but no restriction is imposed on possible recursion. Schabes observes that TSGs are strictly context-free in their weak generative capacity, though the tree sets they produce are somewhat richer.[2]

Suppose now that we wish to generate sentences containing modification structures, e.g. relativization and adverbial adjuncts. It is indeed possible to use TSG to produce linguistically natural derived structural representations. We may however have independent motivations for what may constitute an elementary tree in our grammar. In particular, we might propose (following, among others, Frank 1992) that the elementary trees of a tree rewriting system should contain only information concerning a single predicate, such as a verb, and its associated argument structure. If this is true, then there can be no representation of a modification structure (such as an adverbial modifier) in an elementary tree since it does not play a role in the argument structure of the predicate heading that tree. Consequently, TSG will not be sufficient for our purposes. Instead, we must turn to a somewhat more powerful system of tree rewriting which allows the operation of adjoining, namely Tree Adjoining Grammar (TAG – Joshi, Levy and Takahashi 1975). Adjoining allows us to introduce modification structures into elementary trees which previously lacked any representation of them. Thus, TAG provides us with a richer and more linguistically appealing class of derivations for a set of trees that are generable by TSGs. Note that the weak generative capacity of TAG is strictly greater than the context-free power of TSG, though it is nonetheless restricted to the class of so-called "mildly context-sensitive languages" (Joshi, Vijay-Shanker and Weir 1991).

IV. To summarize, we have seen that there is an increase in generative complexity associated with the tree rewriting systems necessary for coordination, complementation and modification. I suggest that it is precisely this increase in complexity that gives rise to this acquisitional sequence. It is important to observe that we are crucially dealing with complexity measures in terms of tree rewriting systems here rather than the more traditional string rewriting systems. In building syntactic representations, the core problem is the recovery and appropriate factorization of dependencies. These problems are most naturally addressed, I would argue in a tree rewriting framework. Further, from the perspective of string language complexity, both coordination and (right branching) complementation produce regular sets. Consequently, on the basis of a string complexity measure, we would not (contrary to fact) predict any difference in acquisitional difficulty between these cases. In previous work (Frank 1992), I found that other complex sentences structures which are tied to the use of the adjoining operation, but which are at least a priori distinct from modification, show similar delayed acquisition. This provides, I claim, independent confirmation that our tree rewriting based approach is on the right track.

---

[2]We point out that the addition of the device of schematic trees to TSG or to TAG does not increase the weak generative capacity since such node expansion can be simulated using substitution or adjoining, though strong generative capaciy is affected since tree structures of arbitrary arity cannot be generated without the use of such schemas. We leave open the issue of whether children's grammatical systems (as well as those of adults) include such schema all the way through development.

Finally, we can ask whether this limited ability to manipulate systems of tree rewriting is tied to computational load, following the proposals of Joshi (1990) and Rambow (1992). Certain experimental evidence suggests that such an approach is correct: children's grammatical difficulties can be alleviated to some degree if extraneous task demands are diminished (cf. Crain and Fodor 1993 for a review). Thus, if more powerful tree rewriting mechanisms are unavailable simply because they demand too much of the child's resources, they might become available when other resources are freed.

References

Bloom, Lois, Margaret Lahey, Lois Hood, Karin Lifter, and Kathleen Fiess. 1980. Complex sentences: Acquisition of syntactic connectives and the semantic relations they encode. *Journal of Child Language* 7:235–261.

Crain, Stephen and Janet D. Fodor. 1993. Competence and performance in child language. In E. Dromi (Ed.), *Language and cognition: A developmental perspective*. Ablex.

Frank, Robert. 1992. *Syntactic Locality and Tree Adjoining Grammar: Grammatical, Acquisition and Processing Perspectives*. PhD thesis, University of Pennsylvania.

Goodluck, Helen. 1981. Children's grammar of complement-subject interpretation. In S. Tavakolian (Ed.), *Language Acquisition and Linguistic Theory*. MIT Press.

Hsu, Jennifer Ryan, Helen Smith Cairns, and Robert Fiengo. 1985. The development of grammars underlying children's interpretation of complex sentences. *Cognition* 20.

Joshi, Aravind K. 1990. Processing crossed and nested dependencies: an automaton perspective on the psycholinguistic results. *Language and Cognitive Processes* 5.

Joshi, Aravind K., Leon Levy, and Masako Takahashi. 1975. Tree adjunct grammars. *Journal of the Computer and System Sciences* 10.

Joshi, Aravind K., K. Vijay-Shanker, and David Weir. 1991. The convergence of mildly context-sensitive grammatical formalisms. In P. Sells, S. Shieber, and T. Wasow (Eds.), *Foundational Issues in Natural Language Processing*. Cambridge, MA: MIT Press.

McDaniel, Dana and Helen Smith Cairns. 1990. Processing and acquisition of control structures by young children. In L. Frazier and J. de Villiers (Eds.), *Language Processing and Language Acquisition*. Kluwer.

Rambow, Owen. 1992. A linguistic and computational analysis of the third construction. In *Proceedings of the 30th Annual Meeting of the Association for Computational Linguistics*, Newark, DE.

Schabes, Yves. 1990. *Mathematical and Computational Aspects of Lexicalized Grammars*. PhD thesis, University of Pennsylvania.

Tavakolian, Susan. 1981. The conjoined clause analysis of relative clauses. In S. Tavakolian (Ed.), *Language Acquisition and Linguistic Theory*. MIT Press.

Weir, David. 1987. From context-free grammars to tree adjoining grammars and beyond. Technical Report MS-CIS-87-42 (LINC Lab 65), Departent of Computer and Information Sciences, University of Pennsyvania. Dissertation Proposal.